Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

Ontology-constrained multi-LLM scoring of hypothesis support in the predictive processing literature

Hamed Nejat

hamed.nejat@vanderbilt.edu

Vanderbilt University

Alexander Maier

Vanderbilt University

Jesse Spencer-Smith

Vanderbilt University

André M. Bastos

Vanderbilt University

Article

Keywords: predictive coding, large language models, literature synthesis, computational neuroscience, ontology, hypothesis mapping

Posted Date: June 2nd, 2026

DOI: https://doi.org/10.21203/rs.3.rs-9861775/v1

License: This work is licensed under a Creative Commons Attribution 4.0 International License.
Read Full License

Additional Declarations: No competing interests reported.

Ontology-constrained multi-LLM scoring of hypothesis support in the predictive processing literature

Authors: Hamed Nejat[1,*], Alexander Maier[1], Jesse Spencer-Smith[2], André M. Bastos[1]

Affiliations

1. Department of Psychology and Vanderbilt Brain Institute, Vanderbilt University
2. Data Science Institute, Vanderbilt University

[*]Corresponding author: Hamed Nejat, hamed.nejat@vanderbilt.edu.

Abstract

Fragmentation is common in interdisciplinary fields with diverse methods and theoretical commitments. Predictive coding neuroscience is a clear example: its literature spans computational theory, electrophysiology, imaging, behavior, and modeling, creating a synthesis problem that conventional meta-analysis cannot easily resolve. Here, we describe a local multi-LLM pipeline for ontology-constrained literature synthesis. The pipeline reads papers, extracts evidence, incorporates figure descriptions, assembles constrained prompts, and validates outputs against an expert glossary. We defined a predictive-coding glossary of thirty-six concepts grouped into three hypotheses: predictive suppression, feedforward error propagation, and ubiquity. A council of ten local language models scored 31 studies across local and global oddball contexts. This enabled pairwise study-agreement analysis, cross-model comparison, and three-dimensional hypothesis-space mapping. Agreement was stronger for local than global oddball contexts, particularly for feedforward error propagation. Hypothesis-space temperature was lower for local oddball contexts and higher for global oddball contexts, indicating greater dispersion in the latter. These results show that local multi-LLM councils can produce auditable disagreement measurements that map heterogeneous literatures into quantitative evidence spaces. **Keywords:** predictive coding; large language models; literature synthesis; computational neuroscience; ontology; hypothesis mapping

1. Introduction

1.1. The Fragmentation Problem in Predictive Coding Neuroscience

Scientific frameworks rarely develop through empirical support alone[1]. They also face growing theoretical and methodological complexity, becoming large, specialized, and heterogeneous enough that readers may struggle to sustain the full argumentative structure[2–5]. This "fragmentation" of the scientific literature is acute in the domain of computational neuroscience. In particular, mathematical formalism, electrophysiology, behavioral observation, statistical modeling, and philosophical foundations are frequently braided together within a single publication, yet the resulting constructs are not consistently tested at equivalent levels of description. The paradigm of predictive coding serves as an illustrative exemplar of this epistemological phenomenon. Over the past few decades, predictive coding has evolved from a narrow computational account of extra-classical receptive field effects in the visual cortex[6] and the retina[7] into a sweeping, nearly all-encompassing explanatory framework addressing perception[8, 9], learning[10], motor action[11, 12], cortical hierarchy[13, 14], psychiatric disorders[15, 16], and active sensing[11, 17].

As the predictive coding framework has broadened its theoretical scope, the consensus on

what constitutes rigorous empirical support has been put into question[18, 19]. At one end of the theoretical spectrum, predictive coding represents a formal, largely abstract principle of hierarchical inference: higher cortical areas issue generative predictions, lower areas compute sensory residuals or prediction errors, and these errors propagate forward up the cortical hierarchy to continuously update internal models[20, 21]. At another theoretical extreme, predictive coding is treated as a highly specific, falsifiable physiological hypothesis mapping directly onto canonical cortical layers, specific cell classes, distinct oscillatory bands, and directed laminar interactions[14, 22, 23]. For instance, some physiological accounts explicitly associate feedforward prediction error signaling with superficial pyramidal populations communicating via high-frequency gamma rhythms[24], while feedback predictions are associated with deep pyramidal populations communicating via lower-frequency alpha and beta rhythms[25, 26]. Furthermore, the terminology itself is used somewhat inconsistently across the literature. Terms such as prediction, expectation, anticipation, surprise, mismatch, deviance, novelty, and error can be deployed in overlapping but distinct contexts depending on the subfield. For example, a reduction in neural activity upon repeated stimulus presentation might be categorized as prediction suppression, sensory adaptation, or attentional gating, depending on the context.

This ambiguity is magnified by the methodological diversity present in modern neuroscience. Predictive coding claims have been investigated using disparate tools ranging from event-related potentials[27] and functional magnetic resonance imaging univariate contrasts[28, 29] to multivariate decoding[30–32], laminar probe recordings[24, 33, 34], calcium imaging[35–37], and a diversity of behavioral tasks[8, 24, 38], including studies using no-report paradigms[34, 36, 39]. Because human experts read this literature through specialized methodological lenses, traditional meta-analytic synthesis across these subfields does not always scale and is difficult to audit. A case in point is a recent community experiment which performed a large-scale survey of the predictive processing literature[40]. The result involved over one year of intellectual labor, weekly discussions amongst dozens of neuroscientists, and iterated writing and editing, which proceeded as shown in the flowchart of Fig. 1A. This process was valuable and produced consensus amongst the co-authors, while also illustrating the substantial coordination required for manual synthesis. The resulting manuscript was 141 pages long, providing an extensive qualitative synthesis rather than a compact quantitative map of the literature[40]. Moreover, there is currently no widely established mechanism for how to incorporate an increasingly large number of new results into this and other established literature syntheses. We propose that this heterogeneity and the volume of scientific publication create a bottleneck for cumulative theory-building. In the following, we demonstrate an AI-assisted approach designed to address this problem.

1.2. Integrating Large Language Models into Scientific Synthesis

The growth of scientific publication volume motivates automated systems capable of reading, comparing, and structuring literature at large scales (Fig. 1B). AI models can supplement established paradigms for literature synthesis and quantitative meta-analysis, while preserving the value of expert interpretation and conventional meta-analytic methods.

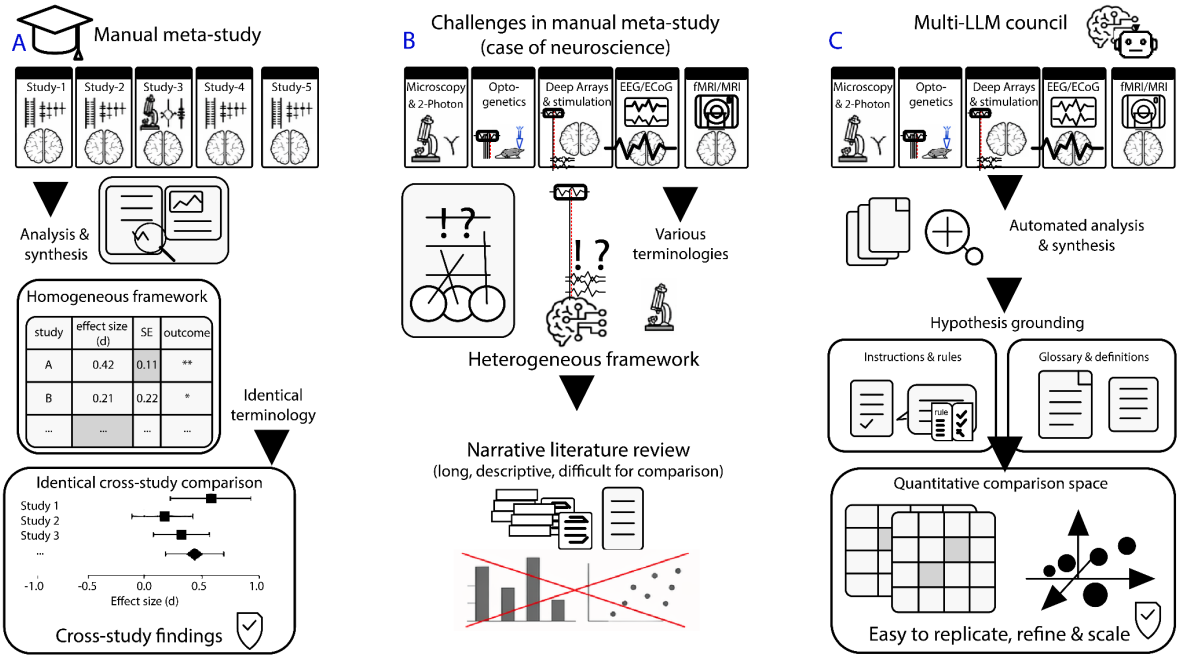


Figure 1: A. Manual meta-studies (e.g., literature review, meta-analysis) are time-intensive and become harder to scale as the magnitude of the literature grows. B. The challenges with a traditional meta-analysis from the perspective of systems neuroscience. Heterogeneous data sources complicate unified statistical treatment. C. The LLM-assisted pipeline we propose here. AI-assisted literature review can ingest a large number of papers and score them on consistent criteria, which can be adjusted by experts. Multiple agents can be run in parallel in a scalable architecture. Replicable output scores can be quantified and compared between studies and agents.

First, in a traditional meta-analysis, papers are generally selected based on a literature which used the same techniques and the same paradigm, and therefore which exist within an identical comparison space (Figure 1A, for a recent application of this approach to the neurobiology of schizophrenia, see Mulvey, Gabhart, et al.[41]). Under these constraints, each study's effect size and reliability is quantified and the consistency of the overall effect across studies is quantified with standard statistics. However, there are growing challenges to this traditional meta-analysis framework. This approach is difficult to apply to heterogeneous literatures where papers use different methodologies and experimental paradigms (Figure 1B). In such settings, there may be no established comparison space and no single standard statistical approach for quantitatively characterizing disparate results (Figure 1B). The common alternative is qualitative literature review, which is valuable but can be difficult to replicate and may not always compress information into comparable variables (for a recent case study in this, see[40]).

To address these challenges, we propose an expert-guided AI-assisted pipeline (Figure 1C) similar to other recent work showing that LLMs can automate multiple aspects of systematic reviews[42] and meta-analyses[43]. The pipeline we propose can ingest and process a large number of heterogeneous papers in parallel and create a structured comparison space based on a canonical, expert-guided glossary. Based on this glossary, a council of multiple independent LLMs can quantitatively score papers on their agreement to a well-defined hypothesis space. Subsequently, the pipeline performs consensus mapping of papers within the hypothesis space. This effectively compresses a large amount of information from heterogeneous sources into a low dimensional

structure of quantitative variables. The output is a geometric constellation, or object, that represents the scientific literature in relation to theoretical predictions that can be further analyzed. New papers can easily be added and compared. Movement of the constellation itself can be tracked over time as a literature evolves, reflecting growing agreement or disagreement with the hypothesis space.

To achieve these goals and to create the proposed pipeline (Figure 1C), it is important to evaluate recent progress of LLMs in scientific reasoning. Indeed, large language models have recently demonstrated extraordinary capabilities across diverse domains, yet their application in rigorous scientific synthesis remains an active, highly debated frontier[5, 44–46].

Within the domain of literature synthesis a recent study (*BrainBench*) showed that large language models tuned to neuroscience (e.g., *BrainGPT*) can forecast neuroscientific results at a level rivaling or exceeding human domain experts[5]. The *BrainGPT* study demonstrated that by capturing the fundamental patterning of methods and results underlying the structure of neuroscience, predictive models can determine whether the scientific content of new abstracts were altered or not more reliably than humans at various levels of domain expertise[5].

Similarly, the emergence of autonomous research frameworks like *The AI Scientist*[45] have illustrated that multi-agent architectures can navigate the entire research lifecycle, from initial ideation and hypothesis generation to coding, result analysis, and even automated peer review (but see[47]). The automated reviewer component of *The AI Scientist*, which operates on an ensemble methodology to evaluate manuscript quality against established conference guidelines, proved capable of predicting paper acceptance decisions with an accuracy on par with human reviewers. The system also generated machine learning results that were accepted in a peer-reviewed conference[45].

More recently, several systems have moved toward structured scientific workflows. Empirical Research Assistance (ERA)[48] coupled LLM-guided code rewriting with tree search to optimize empirical scientific software against explicit, machine-scored objectives, outperforming single LLM calls, best-of-1000 sampling, and AIDE on Kaggle benchmarks. Robin[86] extended multi-agent systems into a lab-in-the-loop biological discovery setting by integrating literature-search agents with data-analysis agents to generate hypotheses, propose experiments, and interpret experimental results. Co-Scientist[87], similarly used an asynchronous multi-agent architecture, under expert supervision, to generate, critique, rank, and evolve scientific hypotheses. Together, these systems suggest that LLMs can contribute to scientific reasoning when constrained by external objectives, specialized agent roles, iterative feedback, and human or experimental validation. These systems also motivate a complementary problem: before autonomous systems generate or test new hypotheses, auditable methods are needed for mapping what the existing literature already supports, contradicts, or leaves unresolved. Here we propose that deploying a constrained, multi-model council equipped with expert-guided ontological guardrails offers a robust mechanism for parsing a fragmented scientific literature like predictive coding.

1.3. Concrete Use Case: Oddball Paradigms in Predictive Coding

Here we examine a conceptually clean yet empirically nontrivial distinction within predictive processing: the separation of the local-oddball (LO) and global-oddball (GO) contexts, a paradigm which has been extensively studied and which has been considered as foundational evidence for predictive processing models in neuroscience[49]. The local-global oddball paradigm was explicitly developed to differentiate short-timescale sensory regularity violations from longer-timescale, abstract rule, or sequence violations[29]. In the local condition, sensory deviance is defined relative

to an immediate, highly probable sensory context, such as a single deviant tone following a rapid succession of standard tones (e.g., xxxY, where "x" and "Y" represent distinct tone frequencies presented to subjects within a few hundred milliseconds). In the global condition, deviance is mapped against a higher-order temporal or abstract sequence expectation (e.g., xxxY xxxY xxxX), requiring the brain to maintain working memory of complex rules across extended temporal windows.

If the predictive coding framework operates as a unified, scale-invariant hierarchical architecture, the generative logic governing the local oddball response should extend to the global oddball response: Higher-level rule violations should generate prediction errors that propagate forward using the same basic inferential mechanics as local sensory violations[49]. However, the empirical literature is still unclear on this point[33, 34, 49, 50]. Divergent neural signatures[34], varying attentional dependencies[29], unique conscious-state sensitivities[33, 51], and differing spatial cortical distributions[29, 34, 52] all complicate a direct one-to-one mapping between local and global contexts. This empirical tension makes the local-global paradigm a useful benchmark for evaluating scientific literature on predictive processing using LLM assistance.

1.4. Rationale for a Local Multi-LLM Council Over a Single Model

A single large language model (LLM) can provide a useful literature synthesis, but a multi-model council offers several methodological advantages. First, model outputs can reflect model-specific priors and blind spots[53, 54]. Second, a single output provides limited information about internal disagreement. In contrast, deploying a benchmarking "council"[55] composed of multiple local language models makes disagreement measurable. Third, multiple agents can help identify unsupported assertions because pooling outputs across agents reduces the influence of uncorrelated model-specific errors[46]. Fourth, multi-model ensembles can improve performance relative to a single model in benchmark settings such as medical question answering[56]. Given these advantages, we used multiple LLM agents to independently score each paper.

Deploying local models provides additional practical and governance advantages. Local execution ensures that the entire analytical process can be accessed and audited. Intermediate markdown files, vision-language model figure extractions, prompt templates, validation statuses, and reasoning logs remain inspectable on local hardware. Furthermore, this approach avoids exposing proprietary scientific documents to external cloud application programming interfaces, ensuring strict data governance, traceability, and reproducibility with fixed review processes, which are paramount for the scientific integrity of automated literature reviews.

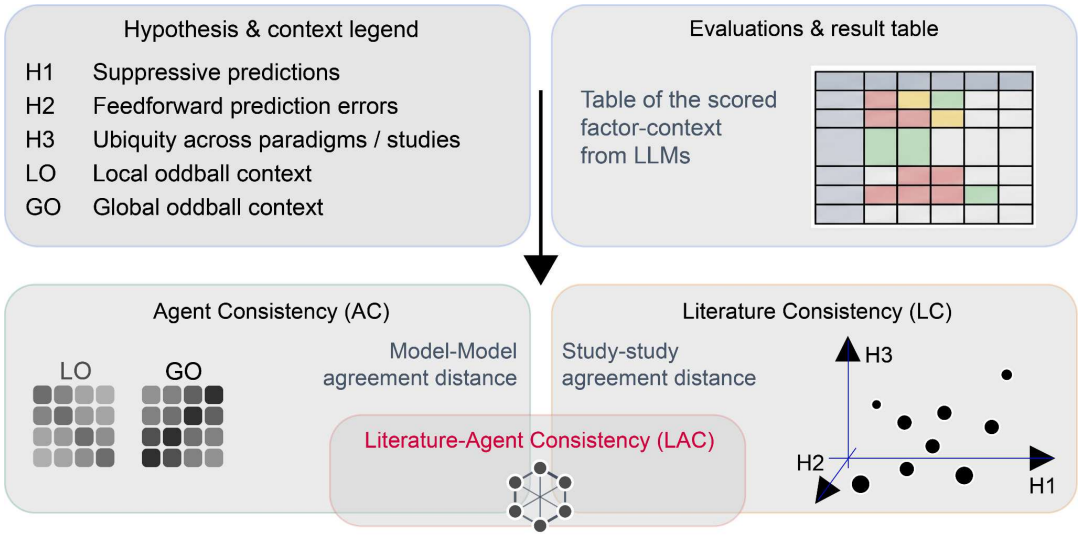


Figure 2: Benchmark overview. A 36-factor glossary defined the Hypothesis space (see Supplementary Tables S1-S3). Two separate contexts are defined, local and global oddballs. The corpus of papers (see Supplementary Table S5) were scored for agreement / disagreement with each hypothesis and in both contexts. Each agent performed these evaluations and saved the structured output in a table. Quantitative analysis of this table defines three benchmarks: *Agent Consistency* (AC, equation 1.1, 1.2) measures how similar different LLMs were to each other in their scores. *Literature Consistency* (LC, equation 2.1, 2.2) measures how similar different papers were to each other in their scores on each of the three Hypotheses. The distance between each paper's position in this 3D space represents a metric of disagreement between any two sets of findings in the literature. *Literature-Agent Consistency* (LAC, equations 3.1-3.3) quantifies a combination of across-agent and across-paper variability.

2. Results

2.1. Agreement Across Hypotheses and Contexts

The local language model council successfully transformed the heterogeneous predictive coding literature into a structured, measurable evidence space. The aggregated results shown in Figure 3 revealed that model evaluations are nonuniform and depend on both the hypotheses (Hypothesis 1, 2, and 3) and the context scope (local and global oddball). The models interpreted most papers as agreeing with these three hypotheses, which form a foundation of the predictive coding literature[10, 14, 59].

Within the local oddball context, the highest score was achieved by evaluations of Hypothesis 2 (Feedforward Error Propagation, Mean = 0.51 ± 0.34, SD). The next highest score was Hypothesis 1 (Predictive Suppression, Mean = 0.46 ± 0.42, SD), followed by Hypothesis 3 (Ubiquity, Mean = 0.31 ± 0.39, SD). The convergence on H1 and H2 suggested that the evidence of suppression-like effects of prediction — such as response dampening or repetition suppression — were consistently described in the literature and had a feedforward pattern of activation in cortical brain areas. Furthermore, these consistently high scores were identified by the diverse models that comprise the council. In contrast, the ubiquity hypothesis (H3, across both contexts) displayed the lowest agreement (Mean = 0.23 ± 0.40), which can be interpreted to indicate that models identified less evidence (compared to H1 and H2, across both contexts) for these prediction error signals to

generalize across the brain.

An assumption within hierarchical predictive coding framework is that global contextual prediction inherits the computational machinery of local sensory prediction, operating on similar predictive coding principles albeit at longer timescales[14, 69, 80]. The global oddball context was also positive for H1-H3, but the scores were lower than the local oddball context for all three hypotheses. There was an average global-to-local downward shift of -0.049 for H1 (Predictive Suppression, $p < 0.05$), -0.158 for H2 (Feedforward Error Propagation, $p < 0.01$), and -0.032 for H3 (Ubiquity, $p < 0.05$, all comparisons, LO vs. GO, paired model-study t-test). The lowest scores across all hypotheses/context groupings were observed in the global oddball context for the ubiquity hypothesis (H3). Across the thirty-one-study corpus, the council consistently recorded lower scores in global oddball contexts.

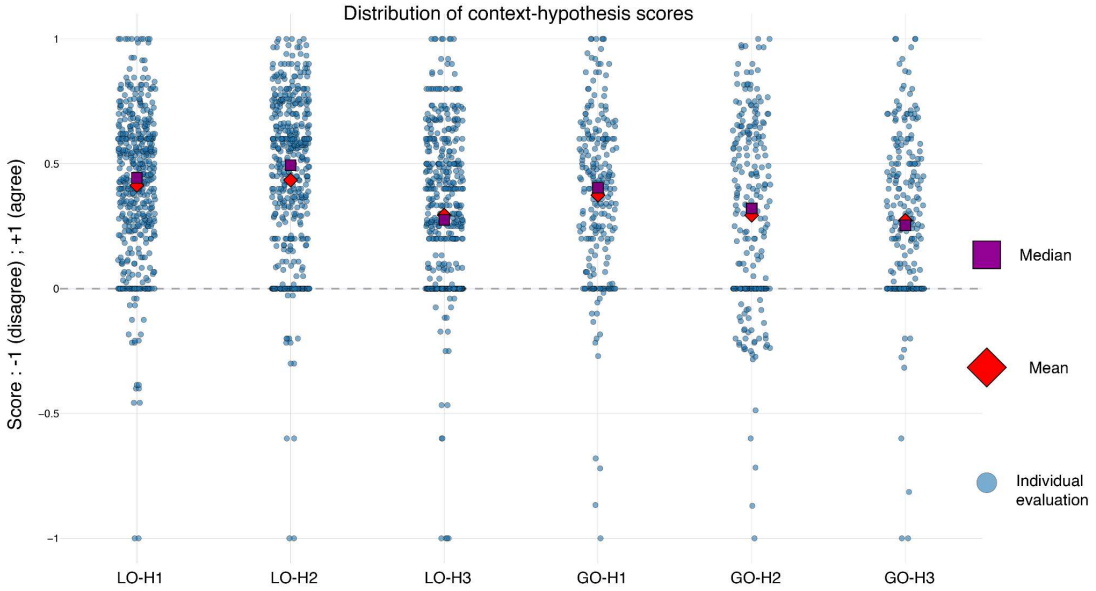


Figure 3: The evaluation result of all papers (N=31) by all models (N=10). Scores are separated by context (LO = Local Oddball, GO = Global Oddball) and by Hypothesis (Supplementary Tables S1-S3, Hypothesis 1: Predictive Suppression; Hypothesis 2: Feedforward Error Propagation; Hypothesis 3: Ubiquity). Each dot is a separate score (total count = 1860). The mean and median of the scores are significantly positive ($p < 0.01$), which is consistent with our choice of these hypotheses as central to predictive coding.

To quantify inter-model agreement for both the local and global oddball context and separately for H1-H3, we used the *Mean Squared Difference* as a distance metric (see Methods, Agent Consistency, equation 1.1, 1.2). Lower values indicate more similar values and therefore stronger agreement/consensus among the models. Agreement between models (Figure 4, AC) had a mean of 0.11. This was well below what could be expected from a random shuffle (0.67) or a hypothesis shuffle (0.5) (for the definition of these shuffle tests, see Methods section, definitions 1.8 and 1.9).

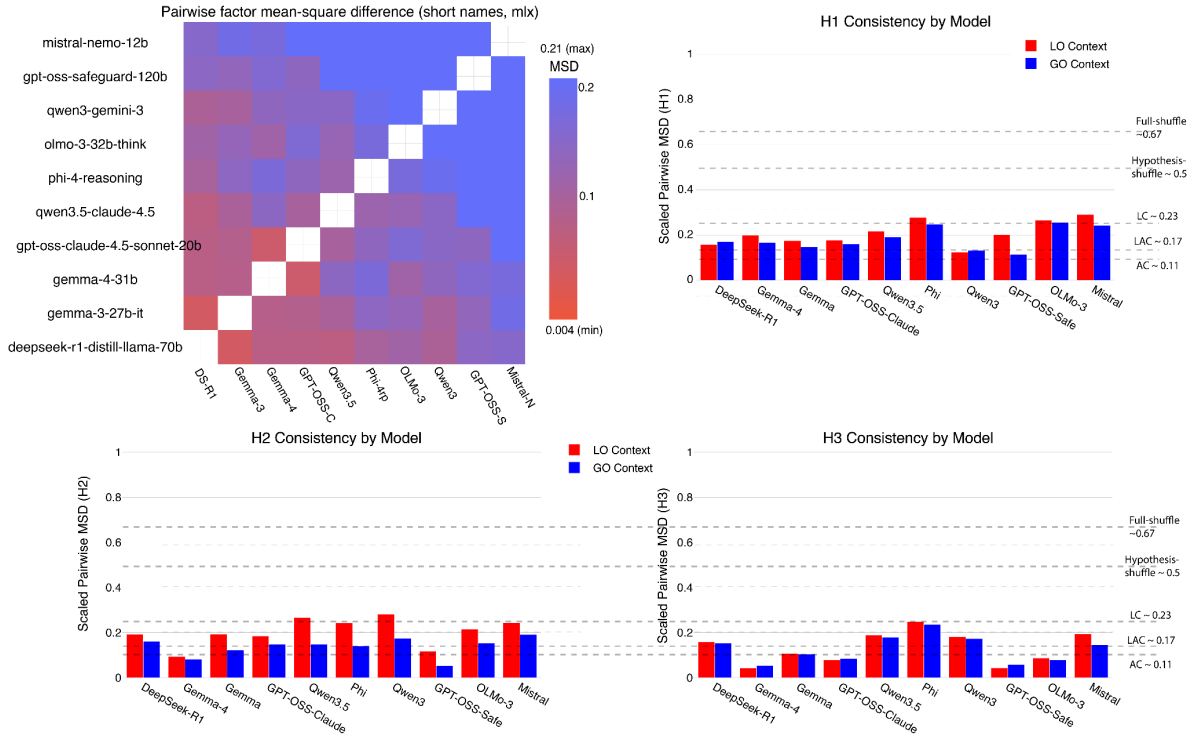


Figure 4: Agent consistency measured as the mean squared distance between the aggregated scores of each of the 10 agents. Agent to agent distance is shown on the upper left. The mean in this comparison was 0.11. Mistral-nemo-12b was the model with the largest distance to the others. The mean squared distance of each model by context is shown for H1 (upper right), H2 (lower left), and H3 (lower right). The dotted lines indicate the MSD distance for a random shuffle of all scores (Full-shuffle, 0.67, def 1.8) and the Hypotheses shuffle (0.5, see Methods, def.1.9). For comparison, values of AC, LC, and LAC are also depicted (see Methods for their respective definitions).

The *Literature Consistency (LC)* benchmark enabled us to quantitatively identify how variable scores were between the papers we considered (Figure 5). Studies such as Bastos et al., 2012, Friston et al., 2010, and Spratling, 2008 exhibited low mean distances (0.07 ± 0.01 SEM), marking them as central reference points that closely align with the dominant trends of the literature. Conversely, Westerberg, Xiong, et al., (2025) emerged as an outlier, with a mean squared difference of 0.464 relative to the rest of the literature. This is consistent with the formulation and conclusions of Westerberg, Xiong, et al., (2025), which diverge from several predictive coding hypotheses, especially in the global oddball context. The LC benchmark therefore enables a geometric mapping between papers and identifies publications that are central to, or divergent from, the established hypotheses.

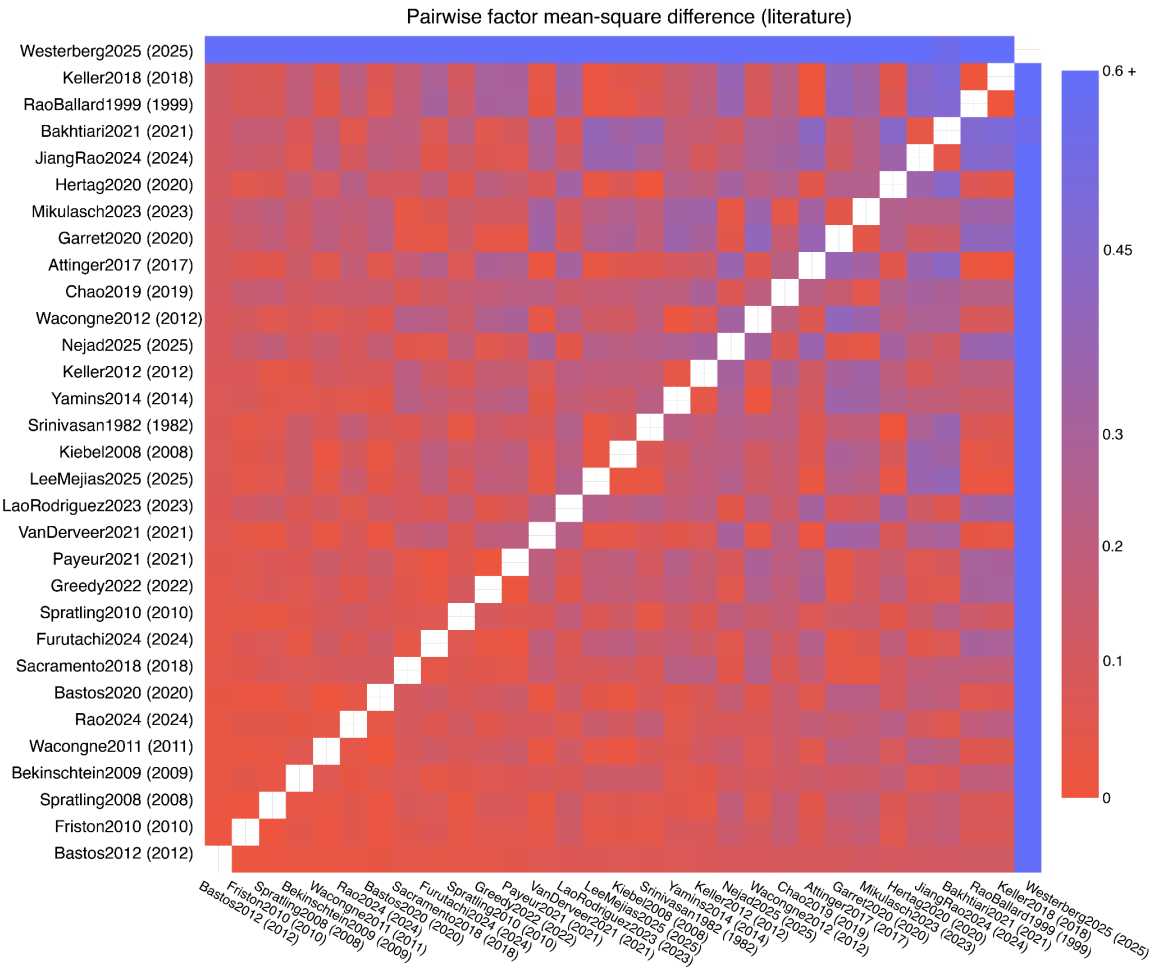


Figure 5: Literature consistency (LC) visualization. The mean squared distance between every paper in the corpus is depicted. Warmer colors indicate lower MSD values or more agreement between studies. The highest values were observed for an outlier paper, Westerberg et al., 2025.

2.2. Direct Comparison of Local and Global Contexts between papers

Further analysis of the distance scores between Westerberg, Xiong, et al., (2025) and the remaining literature revealed an interesting pattern: the scores were significantly greater for the global oddball ($p < 0.01$, mean = 0.55) as compared to the local oddball context (mean = 0.15), see Figure 6. This indicates that the council benchmark captured the pattern reported by Westerberg et al.[34]: stronger divergence from predictive-coding expectations in global oddball findings than in local oddball findings.

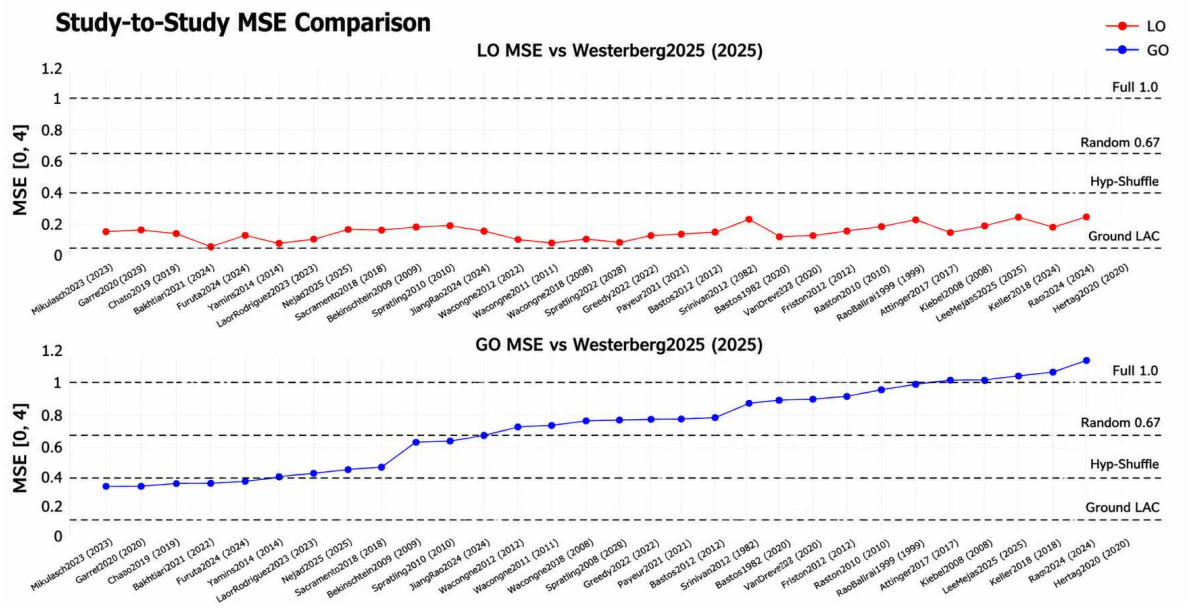


Figure 6: Mean squared error values measuring the distance between each paper and the outlier paper (Westerberg, Xiong, et al., 2025), separately for the local oddball (upper subpanel) and global oddball (lower subpanel) contexts. Papers are ordered from left to right in order of increasing global-oddball MSD to Westerberg, Xiong, et al. (2025).

2.3. Three Dimensional Hypothesis Space Geometry Reveals a Constellation

To visualize the latent structure of agreement and disagreement within the literature, the studies were geometrically embedded into a three-dimensional hypothesis space defined by their H1, H2, and H3 scores, when rendering the local and global contexts separately. The central coordinate of each study reflects the average score separately per hypothesis and averaged across the council of 10 models. Error bars denote the standard error of the mean across the council. Within the local oddball coordinate space, the majority of the studies cluster into a moderately compact, centralized region along the H1, H2 and H3 axes, visually representing the foundational agreement the council reached regarding overall agreement on these hypotheses. For local oddballs, the one exception was Westerberg et al. (2025), which had positive scores on H2 and H3, but a moderately negative score on H1 (Predictive Suppression).

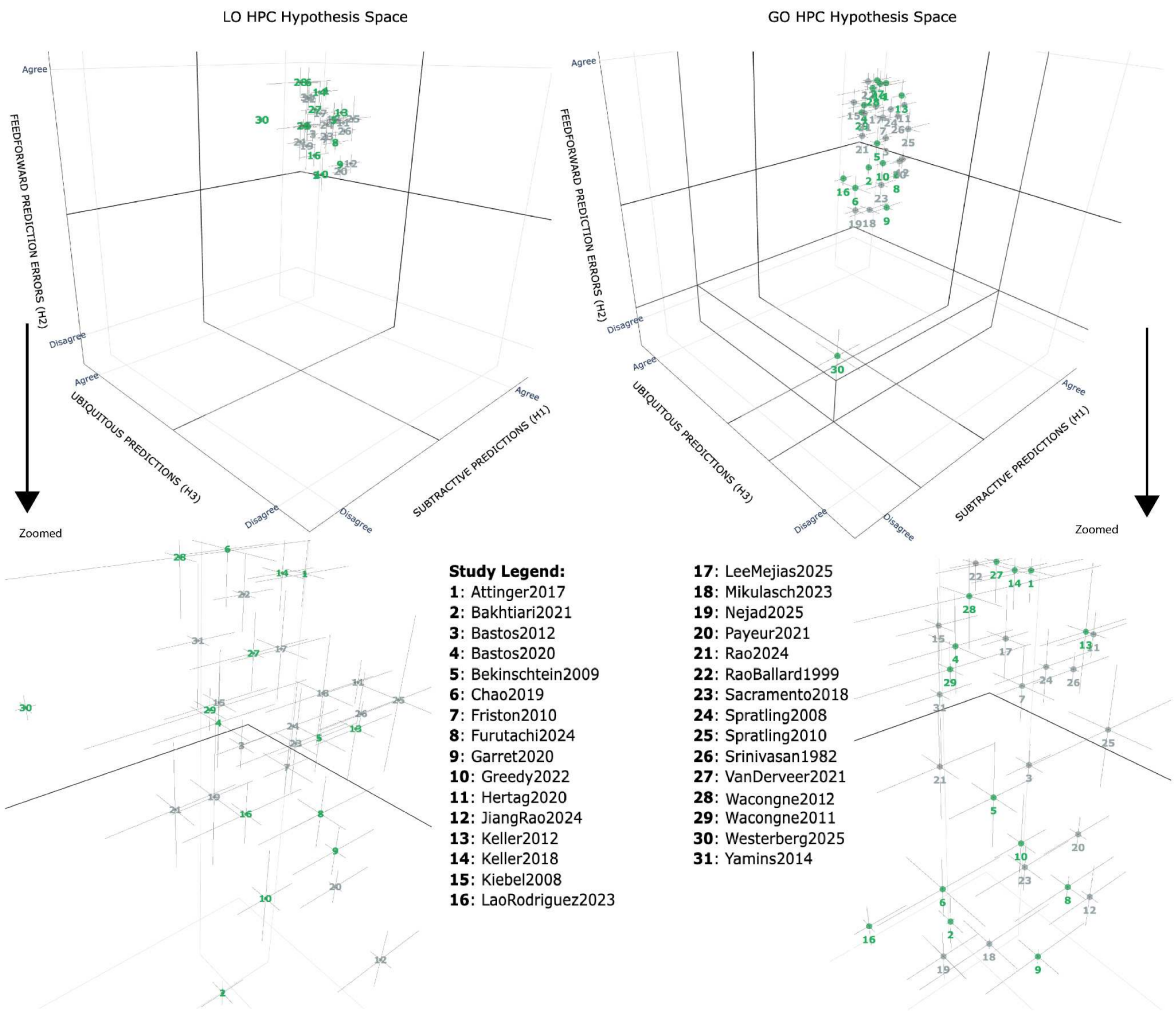


Figure 7: Three-Dimensional view of the scores of each paper, with scores shown on the H1 (Predictive Suppression), H2 (Feedforward Prediction Errors), and H3 (Ubiquity) axes. The 3 lines emanating from each of the studies reflects the standard error of the mean of those scores across the 10 evaluator models. The local oddball context is shown on the left plot and the global oddball context is shown on the right.

Moving from local oddball (Figure 7, left) to global oddball (Figure 7, right) there was a significant move away from the +1, +1, +1 octant, with multiple studies going negative on H2. Study 30 (Westerberg, Xiong, et al., 2025) is the outlier, as it had the strongest movement between Local and Global Oddball contexts (Figures 7 and 8). The move from Local to Global Oddball contexts created a literature that was more dispersed or variable.

To quantify this shift, we introduced a novel metric of dispersion or variability within a well-specified hypothesis space. Inspired by physics, we call this metric the hypothesis-space-temperature. We define the hypothesis-space-temperature by fitting an ellipsoid that minimizes the volume required to fit all points in three dimensions, divided by the number of data points (studies) that are inside of the ellipsoid (see equation 3.4):

Hypothesis space temperature $T_H = c\,V_{\text{env}}/n, \quad c = 1.0$

The scaling constant c is fixed to 1.0 for this analysis. Here, T_H denotes the metaphorical hypothesis-space temperature, V_{env} is the volume of the minimum volume enclosing ellipsoid containing all hypothesis points, and n is the number of points. (eqn. 3.4)

Operationally, when a hypothesis space provides a compact description of a literature, a new observation is expected to fall within the existing ellipsoid. In that case, the ellipsoid volume can remain approximately stable while the number of samples increases, thereby decreasing temperature. In contrast, a new study that falls outside the existing ellipsoid can increase the ellipsoid volume, increasing temperature in proportion to V/n (see equation above).

We calculated hypothesis-space-temperature separately for the Local and Global oddball contexts, and extracted the standard error of the mean of this measure across different LLM models. The hypothesis-space-temperature of Local Oddballs (excluding Westerberg, Xiong et al., 2025) was 0.00114 (SEM across N=10 models, 0.00016, $p < 0.01$), significantly lower than for Global Oddballs, 0.00348 (SEM across N=10 models, 0.00039, $p < 0.01$). This confirmed what can be appreciated visually in Figure 7, namely that the dispersion of studies grows larger in the Global Oddball compared to Local Oddball contexts, even without the outlier study. This effect was even stronger when including all 31 studies: across all studies, hypothesis-space-temperature of Local Oddballs was 0.00177 (SEM across N=10 models, 0.00033), significantly lower for Global Oddballs, 0.00566 (SEM across N=10 models, 0.00111). This geometric visualization indicates that the predictive coding literature is structured by a dense central neighborhood and distinct, isolated data points that increase dispersion in the hypothesis space.

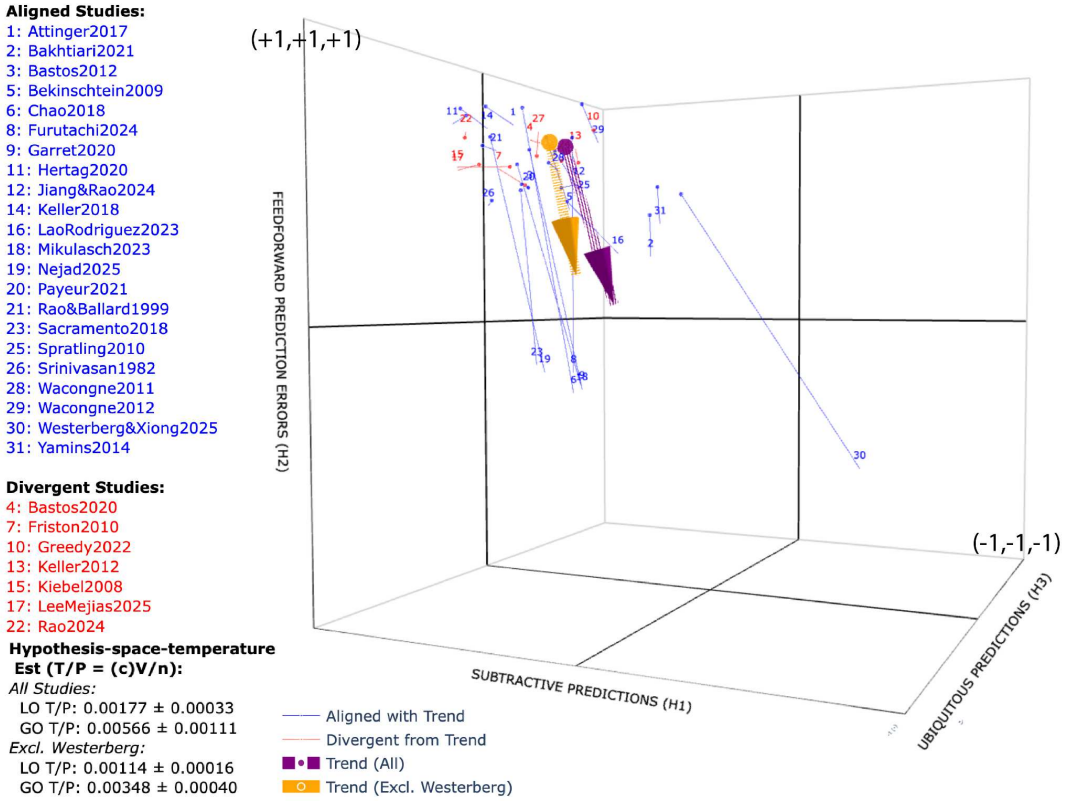


Figure 8: Three-Dimensional vectors of each paper from the Local to the Global oddball position. Each paper's vector is shown independently. The average movement (vector direction) without

the outlier study (Westerberg, Xiong, et al., 2025) is shown in yellow, and the average movement of all studies is shown in purple. Individual study's vectors are labeled as blue if their individual vector was aligned to the Westerberg, Xiong, et al., (2025) study's vector on all three axes, and red otherwise.

In addition to an increase in temperature, it is also apparent in Figure 7 that the change from Local to Global Oddball contexts appears to shift along a particular axis. To quantify this, we calculated a vector for each study between its Local and Global Oddball position. This is shown in Figure 8. Each vector represents an individual study's movement magnitude and direction. The mean vector across all studies is depicted in purple, while the vector of all studies excluding Westerberg, Xiong et al. (2025) is shown in yellow. Both vectors are aligned in a similar direction pointing away from the +1, +1, +1 octant (which would indicate full agreement with all three hypotheses). We interpret this to mean that the literature largely agrees with the predictive coding hypothesis space for Local Oddballs but much less agreement is observed for Global Oddballs. Furthermore, while the Westerberg, Xiong, et al. study is an outlier in the sense that the movement vector magnitude is the largest of all studies, the movement vector angle of this study is aligned to the underlying Local-Global shift within the pre-existing literature.

3. Discussion

Hidden structure within the predictive coding literature.

The application of the multi-model benchmarking council shows that a heterogeneous predictive coding literature can be organized into structured, quantitatively identifiable patterns. The council's outputs showed that the literature can be organized by structured, quantitatively identifiable scores that draw out important distinctions between papers and contexts. One pronounced distinction was identified between stronger local oddball support and lower model-estimated support in the global oddball context. The models' agreement scores for H1, H2, and H3 were all higher for the local as compared to global contexts, which indicates stronger evidence for predictive coding in the local oddball domain. This context-dependent reduction was most pronounced for H2 (Feedforward Error Propagation), indicating that support for this hypothesis varies most strongly between local and global oddball contexts.

Furthermore, formalizing this quantitative hypothesis space permitted the extraction of geometric parameters of the corpus. Mapping the change of each individual study's movement from local to global paradigm with vectors allowed us to calculate theoretical shifts via mean vector displacement. This metric models directional movement between local and global paradigms, highlighting differences in these experimental contexts. Also, we derived a hypothesis-space-temperature metric. High temperature indicates hypothesis state-space volatility and theoretical disagreement. Low temperature signifies structural consensus. This provides the compass for a quantitative cartography of the literature, translating abstract scientific debate into measurable variables.

Scientific synthesis with multi-model councils.

For neuroscience and related interdisciplinary fields, our findings support the utility of multi-agent architectures for complex scientific analyses, echoing other AI-based advances such as *The AI Scientist*[45]. A single language model can summarize these thirty-one papers, but a council provides additional visibility into model-to-model variability, theoretical disagreement, and methodological nuance present in the corpus.

In contrast, the council framework quantifies and utilizes disagreement as an analytical tool. By

calculating *Agent Consistency*, *Literature Consistency*, and *Literature-Agent Consistency* baselines, researchers can map whether a particular scientific result is stable across contexts or context-sensitive and model-specific. In addition, random shuffling of model-produced values defines a statistical reference distribution under random scores. Finally, multiple model scores allow uncertainty over models to be estimated; here we used the standard error of the mean (Figure 7). The council framework therefore treats language-model outputs as structured, auditable scientific measurements when context and instructions are constrained.

Importance of an explicit glossary and ontology.

A conclusion drawn from this study is that, in artificial-intelligence-assisted scientific synthesis, the ontology matters alongside the underlying foundation models. The glossary provides the council with a stable conceptual target. The instruction layer and anti-guess rules[58] further constrain model outputs toward explicit textual evidence rather than model-specific priors. In addition, all scores were accompanied by an auditable reasoning log. Inspection of a subset of these logs (see Section 2.4) revealed domain-appropriate reasoning patterns, while also identifying occasional divergences. These considerations suggest that the quality of language model benchmarking depends on task design, problem framing via a glossary, and the deployment of control layers. The ontology provided by the canonical glossary instructs the models what to look for, and allows us to understand and interpret what the benchmark signifies.

Methodological limitations and future expansions.

This framework has several methodological boundaries that guide future expansion. The corpus size of thirty-one studies, drawn largely from a recent study's comparisons[34], supports a focused initial cartography rather than exhaustive coverage of the predictive coding field. Second, the pipeline relies on the specific thirty-six factors chosen. Designing an alternative ontology based on active inference or variants of predictive coding theories could shift the geometric clustering. Because scores are nested within papers, models, factors, and contexts, future extensions can use hierarchical or mixed-effects models to estimate uncertainty at each level.

Future iterations of the pipeline can expand the corpus to enable higher-resolution analyses across decades of publications and additional subfields. In principle, a seed study could initiate corpus expansion through automated citation and similarity search. Future work can also examine whether ontological definitions of the hypothesis space can be proposed or refined automatically. In that setting, more than three dimensions could be included, and hypothesis dimensions could be added or revised as the literature evolves. This also points to the possibility of measuring literature dynamics over time. As scientific literatures shift, new findings revise established interpretations, and support for hypotheses waxes or wanes, extensions of the present approach could capture and quantify these changes, allowing meta-scientific analyses to guide future studies.

Conclusions. This research establishes a local multi-LLM benchmark framework designed to evaluate the predictive coding literature in neuroscience. By using ten local language models with a grounded glossary, instruction layers, visual extraction, and validation, the pipeline converts heterogeneous scientific papers into a quantifiable evidence space. Within this framework, assistance from large language models can be effective at mapping the empirical support for specific hypotheses that underlie scientific knowledge. The results illustrate that the literature supports consensus regarding local predictive suppression, but exhibits more disagreement concerning broad ubiquity claims and global feedforward error propagation. The integration of metrics such as *Agent Consistency*, *Literature Consistency*, and the *Literature-Agent Consistency* provides a foundation for identifying consensus and empirical outliers. This enables expert-guided, automated, and replicable

data-model comparisons between sets of results, including theoretical, conceptual, empirical, and computational papers. The distance metrics developed here provide a generalizable framework for hypothesis-space comparison. New evidence that differs from existing literature will be distant from the rest. As the weight of this evidence builds with additional datasets and published studies, the constellation of the underlying literature is altered. The methods introduced here can visualize and quantify these shifts. This captures the fundamental process of cumulative scientific reasoning itself, as additional evidence either fortifies existing hypotheses or renders them less likely. Additional hypotheses can also alter the geometric structure of scientific reasoning by adding additional dimensions beyond the three hypothesis space representations we considered here.

4. Methods

The multi-LLM processing pipeline, spanning from the initial ingestion of the raw publications to the visualization of the hypothesis space, was constructed to ensure that structured evidence can be traced directly back to its originating paper, its specific glossary factor, and its model-specific reasoning log.

4.1. Overview of the Ontological Benchmark Pipeline

The workflow was initiated by defining the hierarchical predictive coding framework and building the canonical glossary (Supplementary Tables S1-S3) and instruction layer. Following this, source PDFs (Supplementary Table S5) were ingested through deterministic text parsing. Each of a paper's figures were also ingested via a figure-aware step using *DeepRead*[57]. Once the study text and figure descriptions were consolidated into a unified markdown file, the system assembled a constrained evaluation prompt and dispatched it to the ten independent scoring models (Supplementary Table S6). The resulting outputs were validated and organized (based on the rules section of the input to the LLM) into a comprehensive table, which served as the database for all summary matrices and hypothesis-space visualizations (Supplementary Fig. S1).

4.2. Corpus Structure and Benchmark Matrix

The working corpus (Supplementary Table S5) evaluated in the current study consisted of thirty-one published, preprint, and archived neuroscience studies, with publication years ranging from 1982 to 2025. These studies were selected because they span the heterogeneity of the predictive coding field, and included theoretical contributions, conceptual reviews, neurophysiology, functional imaging, and computational modeling studies. These studies were included in a recent paper as representative of predictive coding models[34]. We recognize that they do not comprehensively represent the entirety of the predictive coding field, but are deemed sufficiently representative to demonstrate the utility of this pipeline. We plan to apply future extensions of this pipeline to a more comprehensive collection of papers, and the present work serves as a starting point. Performing these evaluations on a smaller set of papers also enables manual checking and verification of the results.

4.3. Hierarchical Predictive Coding Glossary

At the core of the benchmark is the thirty-six-factor glossary. Without a highly constrained glossary, large language models can rely on broad autoregressive priors, generating narrative summaries that may blur critical neuroscientific distinctions and the boundaries required for hypothesis testing[58]. Following the logic proposed by Westerberg et al.[34], we defined a glossary organized by three

distinct hypotheses: predictive suppression (H1), feedforward error propagation (H2), and ubiquity (H3). The full glossary, detailing the specific factors, their definitions, and their applicable context scopes, is provided in Supplementary Tables S1-S3.

Hypothesis 1 (Predictive Suppression) encompasses the core phenomenon that initially motivated predictive coding: that expected inputs are suppressed, sharpened, or explained away. Factors within this family (Supplementary Table S1), such as activity suppression or GABAergic inhibition serve to encompass the idea that top-down predictions suppress bottom-up sensory signals[6, 14].

Hypothesis 2 (Feedforward Error Propagation) specifies that unpredicted mismatch is routed and propagated forward through the cortical hierarchy. This requires the models to search for signatures such as supragranular layer involvement[59], gamma-band signaling[14], and directed bottom-up connectivity[24] (Supplementary Table S2).

Hypothesis 3 (Ubiquity) captures one of the most expansive theoretical claims, evaluating whether predictive coding serves as a universal algorithm of the cortex[10, 14]. H3 guides the models to discern whether a study provides evidence for mechanisms that are broadly consistent across-species, modalities, and/or brain areas (Supplementary Table S3).

Each factor was further described with more context regarding quantitative, qualitative, or methodological tags. The glossary dictated that not all textual evidence should be treated equally. A quantitative tag demands that the language model locate explicit statistical results, tables, or graphical data. The context scope column dictates the applicability of the factor to the local and global contexts. Factors directly tied to short-timescale adaptation effects are restricted strictly to local oddball scoring, preventing the benchmark from scoring sensory habituation as evidence for long-timescale sequence learning. The following papers were passed to the pipeline (Supplementary Table S5).

4.4. The Instruction Layer, Prompt Methodology, and Constraints

A glossary alone is likely insufficient if the language model acts as an unconstrained generative engine[58]. Large language models can follow conversational priors, infer missing information, or substitute generic scientific plausibility for explicit textual evidence[82]. Consequently, the pipeline separated the glossary from an instruction layer that specifies how the model evaluates the glossary.

The prompt architecture is initiated by establishing a specific role, instructing the language model to evaluate the paper as a neuroscientist. This role specification was intended to activate domain-relevant representations[58], orienting the model toward complex methodological details and the analytic standards typical of high-level peer review. The prompt then established a continuous scoring methodology for each factor ranging from negative one to positive one (within one decimal point, e.g., -0.2, +0.1). Positive scores indicate empirical support, negative scores indicate explicit contradiction, and precision is encouraged through the use of intermediate values to capture the magnitude of agreement/disagreement, given the evidence.

The prompt emphasized a distinction between scoring a zero and scoring a null value. A score of zero is assigned for instances where the study explicitly addresses a particular factor, but reports statistically neutral results that neither agree nor disagree with a factor. A null value was returned if the study does not address or measure the factor in any capacity (e.g., a study using EEG would not evaluate hypotheses related to cortical layers or specific cell types). Instructing the models to use null for unaddressed factors reduces false negatives and preserves the absence of evidence within

the literature.

The instruction layer is further specified by anti-guess constraints. Models were instructed not to invent novel glossary keys, rename existing factors, or force local-global scoring onto irrelevant papers. Furthermore, the instructions required a reasoning log within the output. For every non-null score assigned, the model was required to cite the specific section, figure, or statistical table from the source text that justified its evaluation. This transformed the output from a standalone numerical assertion into an auditable chain of evidence, which we used to inspect whether unsupported assertions were present. We randomly inspected a subset of logs to confirm this. The full outputs are available on the repository (see *Data and Code Availability*).

To give one instructive example, Mistral-Nemo scored Westerberg et al., 2025, with a reasoning that included "... *I score H2: Feedforward Deviance as positive. In LO (Local-Oddball) Fig 2b, I find hierarchical progression [+0.8]. However, GO (Global-Oddball) signals in Fig 3 show no temporal order of response [-0.6], suggesting an alternative mechanism compared to LO ...* ".

4.5. Deterministic Literature Ingestion

Text extraction alone is insufficient for neuroscience literature, where substantial inferential signal is frequently encoded within a paper's figures. A purely text-based evaluation pipeline would omit this empirical evidence. To address this, the pipeline incorporated a multimodal processing stage. The pipeline then extracted the figures and analyzed them with a local vision-language model (VLM) instructed to operationalize graphical data (gemma-3-9b-it VLM), decode statistical trends, and map visual evidence directly onto the predictive coding glossary. These descriptive insights were subsequently interleaved into the study markdown document which was passed to the LLM models.

4.6. Model Council Orchestration and Validation

The following 10 models comprised the council of LLMs that were used for evaluation (Supplementary Table S6):

The pipeline contained a validation tool designed to audit syntactic integrity, verify top-level key presence, and ensure that factor nomenclature matches the canonical glossary. Any output that did not meet these constraints was flagged for exclusion or revision, preventing misaligned data (e.g., incorrect factor assignment) from entering the downstream hypothesis-space visualizations.

The LLMs in the pipeline were configured during model initialization via mlx-lm package utilizing the same sampling and memory optimization parameters for all models. Inference parameters were set to a sampling temperature of $\tau = 0.70$ with top-p/min-p of $p = 0.9$ and $\alpha = 0.1$ [84] , respectively. To manage computational overhead during long-context execution, the sequence length for each model was strictly bounded by a context window of $L = 2^{17}$ (131,072 tokens), with past attention states preserved in a KV-cache using MXFP8[85] precision to optimize memory footprint and maximize throughput.

4.7. Construction of the Hierarchical Predictive Coding Table and Measures of Evaluation Consistency

Validated outputs were flattened to construct the Hierarchical Predictive Coding Table. This matrix aligned the data into three hundred and ten rows, with each row corresponding to a unique study(s)-model(m) pair. The matrix comprises columns capturing metadata, and factor-level columns representing all thirty-six factors (f) across both local and global contexts (c). The table

enables multidimensional hypothesis visualization and pairwise agreement correlation on the scores. Evaluation scores that were not applicable for a subset of factors were ignored and omitted from calculations.

The table holds values for each study s, model m, context c, and factor f, (def 1.1):

Score Array Element	$S(s, m, c, f) \in [-1, 1] \cup \{\emptyset\}$
Score value ranging from −1 (disagrees) to 1 (agrees), or undefined (∅). (def. 1.1)	

Where each of these s, m, c, and f elements contain the following indices (def 1.2):

Score Indices	$s \in \{1, \ldots, 31\},\ m \in \{1, \ldots, 10\},\ c \in \{\mathrm{LO}, \mathrm{GO}\},\ f \in \{1, \ldots, 36\}$
Indices for study (s), model (m), context (c), and factor (f). (def. 1.2)	

The 36 factors are defined in Supplementary Tables S1-S3 and their indices correspond to Hypothesis 1 (Predictive Suppression), Hypothesis 2 (Feedforward Error Propagation), and Hypothesis 3 (Ubiquity). All evaluations had scores for each hypothesis to provide a non-missing numeric value for agreement with hypotheses, (def 1.3):

Hypothesis Sets	$H_1 = \{1, \ldots, 12\},\ H_2 = \{13, \ldots, 24\},\ H_3 = \{25, \ldots, 36\}$
Partition of factors into three hypotheses. (def. 1.3)	

By averaging across studies (s), models (m), and contexts (c) it is possible to define each score for each hypothesis. Note that in the equation below, null values, which models are instructed to return when it is not possible to evaluate a factor, are removed from the mean (def 1.4):

Mean Hypothesis Score	$\bar{S}_h(s, m, c) = \dfrac{\sum_{f \in H_h} \mathbf{1}[S(s, m, c, f) \neq \emptyset] S(s, m, c, f)}{\sum_{f \in H_h} \mathbf{1}[S(s, m, c, f) \neq \emptyset]}$
Mean score across factors in hypothesis H_h for fixed (s, m, c). (def. 1.4)	

The same Mean Hypothesis score can also be expressed as an average across all 10 models (m) that comprise the council (def 1.5):

Grand Mean Hypothesis Score	$\bar{S}_h(s, c) = \dfrac{1}{10} \sum_{m=1}^{10} \bar{S}_h(s, m, c)$
Mean hypothesis score across all models for a given (s, c). (def. 1.5)	

We calculated pairwise mean-square differences for non-missing subset of the scores across hypotheses and contexts (all of the evaluations had at least 3 factors scored for each set, def 1.6):

Mean-Square Difference (MSD)	$\mathrm{MSD}(A, B) = \dfrac{1}{6} \sum_{c \in \{\mathrm{LO,GO}\}} \sum_{h=1}^{3} \big(A(c, h) - B(c, h)\big)^2$
Squared difference between two hypothesis-score arrays. (def. 1.6)	

To compute the statistical significance of these scores and their distances, we performed random permutation tests, where scores were randomly permuted across the table either across all dimensions (full shuffle) or across a subset of dimensions (e.g., model, context, and factor fixed but randomizing study, def. 1.7, 1.8):

Permutation Operator	$\pi : \Omega \to \Omega$ (def. 1.7)
A random permutation sampled uniformly, $S_{\|\Omega\|}$, representing a discrete shuffle of the index space Ω.	

Full Shuffle	$\tilde{S}(i) = S(\pi(i)), \quad i = (s, m, c, f)$ (def. 1.8)
A global shuffle where π permutes the entire flattened multi-dimensional index space simultaneously.	

For the hypothesis shuffle, we kept the model and the study fixed, but shuffled context and factors of hypotheses. (def. 1.9):

Hypothesis Shuffle	$\tilde{S}(s, m, j) = S(s, m, \pi_{s,m}(j)), \quad j = (c, f)$ (def. 1.9)
A particular permutation applied to the context-factor indices (c, f) for each fixed study and model.	

4.8. Agent Consistency (AC)

Agent Consistency (AC) quantifies pairwise agreement between models, showing the relationship between the models in the council (equation 1.1). AC evaluates how closely individual models align in their textual interpretations across all studies (s):

Agent Consistency (Pairwise)	$\mathrm{AC}(m_i, m_j) = \dfrac{1}{31} \displaystyle\sum_{s=1}^{31} \mathrm{MSD}\big(\bar{S}_h(s, m_i, \cdot), \bar{S}_h(s, m_j, \cdot)\big)$
Agreement between two models across studies. (eqn. 1.1)	

Averaging AC across all possible unique model combinations (for 10 models, 10*9/2=45) yields (equation 1.2), which is plotted as dotted lines in the Figures and indicated as AC:

Grand Agent Consistency	$\mathrm{AC} = \dfrac{1}{45} \displaystyle\sum_{1 \le i < j \le 10} \mathrm{AC}(m_i, m_j)$
Average consistency across all model pairs. (eqn. 1.2)	

4.9. Literature Consistency (LC)

Literature Consistency measures the study-study Mean-Squared Distance across the corpus, calculating how closely different publications align within the hierarchical predictive coding hypothesis space (equation 2.1):

Literature Consistency (Pairwise)	$\mathrm{LC}(s_i, s_j) = \mathrm{MSD}\big(\bar{S}_h(s_i, \cdot), \bar{S}_h(s_j, \cdot)\big)$
Agreement between two studies. (eqn. 2.1)	

Averaging LC across all possible unique study combination (for 31 studies ((31*30)/2)=465) yields (equation 2.2), which is plotted as dotted lines in the Figures and indicated as LC:

Grand Literature Consistency	$\mathrm{LC} = \dfrac{1}{465} \displaystyle\sum_{1 \le i < j \le 31} \mathrm{LC}(s_i, s_j)$
Average consistency across study pairs. (eqn. 2.2)	

4.10. Literature-Agent Consistency (LAC)

The *Literature-Agent Consistency (LAC)* represents a combined consistency score considering both models (m) and studies (s). Since AC,LC and LAC are all distance-based, lower values indicate greater consistency. Calculating LAC with respect to studies yields equation 3.1:

LAC (Literature)	$\mathrm{LAC}_{\mathrm{lit}}(s,m) = \mathrm{MSD}\big(\bar{S}_h(s,m,\cdot), \bar{S}_h^{(-m)}(s,\cdot)\big)$
Model vs. other models for same study. (eqn. 3.1)	

Calculating LAC with respect to models yields equation 3.2:

LAC (Agent)	$\mathrm{LAC}_{\mathrm{agent}}(s,m) = \mathrm{MSD}\big(\bar{S}_h(s,m,\cdot), \bar{S}_h^{(-s)}(m,\cdot)\big)$
Model vs. other studies for same model. (eqn. 3.2)	

Using equation 3.1 and 3.2 to calculate grand LAC, a distance-based scalar value used to represent overall agreement between literature and agents, in equation 3.3:

Grand LAC	$\mathrm{LAC}_{\mathrm{grand}} = \dfrac{1}{2\lvert S\rvert\lvert M\rvert} \displaystyle\sum_{s\in S}\sum_{m\in M} \big[\mathrm{LAC}_{\mathrm{lit}}(s,m) + \mathrm{LAC}_{\mathrm{agent}}(s,m)\big]$
Combined literature-agent consistency aggregated across all studies and models. (eqn. 3.3)	

4.11. Statistical analysis

All summary statistics were computed from the validated model-study-context-factor score table. Scores marked as undefined were excluded from factor-level means rather than treated as zero. Hypothesis-level scores were computed by averaging non-missing factor scores within each hypothesis, then aggregating across models or studies as specified by the equations. Pairwise distances were quantified using mean-square difference over the local and global contexts and the three hypothesis dimensions. Local-versus-global context comparisons were evaluated as paired comparisons over matched study-model units or model-level aggregates, as indicated in the relevant analysis. The manuscript reports nominal alpha thresholds for the comparisons. The analysis code and exported score tables provide the basis for reproducing the reported comparisons, permutation references, and summary statistics.

Data availability

The code, glossary, prompt templates, public-safe configuration placeholders, reproducibility notes, and citation metadata are available at https://github.com/HNXJ/mllm-public. Model outputs and derived score tables required to reproduce the analyses are available from the corresponding author on reasonable request and are planned for archival deposition with the public repository.

Code availability

Analysis and pipeline code are available at https://github.com/HNXJ/mllm-public. The repository includes the HPC-36 predictive-coding glossary, prompt/instruction files, model/runtime documentation, public-safe configuration placeholders, reproducibility notes, and citation metadata.

References

[1] Feyerabend, P. *Against Method: Outline of an Anarchistic Theory of Knowledge*. (Verso, London, 2002).

[2] Asai, A. *et al.* Synthesizing scientific literature with retrieval-augmented language models. *Nature* 650, 857–863 (2026).

[3] Zhang, Y. *et al.* Exploring the role of large language models in the scientific method: from hypothesis to discovery. *npj Artif. Intell.* 1, 14 (2025).

[4] Lu, Y. *et al.* Learning to Generate Structured Output with Schema Reinforcement Learning. (2025) doi:10.48550/ARXIV.2502.18878.

[5] Luo, X. *et al.* Large language models surpass human experts in predicting neuroscience results. *Nat Hum Behav* 9, 305–315 (2024).

[6] Rao, R. P. N. & Ballard, D. H. Predictive coding in the visual cortex: a functional interpretation of some extra-classical receptive-field effects. *nature neuroscience* 2, 79–87 (1999).

[7] Srinivasan, M. V., Laughlin, S. B. & Dubs, A. Predictive coding: a fresh view of inhibition in the retina. *Proc. R. Soc. Lond., B, Biol. Sci.* 216, 427–459 (1982).

[8] Bell, A. H., Summerfield, C., Morin, E. L., Malecek, N. J. & Ungerleider, L. G. Encoding of Stimulus Probability in Macaque Inferior Temporal Cortex. *Current Biology* 26, 2280–2290 (2016).

[9] Issa, E. B., Cadieu, C. F. & DiCarlo, J. J. Neural dynamics at successive stages of the ventral visual stream are consistent with hierarchical error signals. *Elife* 7, (2018).

[10] Friston, K. The free-energy principle: a unified brain theory? *Nature Reviews Neuroscience* 11, 127–138 (2010).

[11] Adams, R. A., Shipp, S. & Friston, K. J. Predictions not commands: active inference in the motor system. *Brain Struct Funct* 218, 611–643 (2013).

[12] Friston, K. What Is Optimal about Motor Control? *Neuron* 72, 488–498 (2011).

[13] Markov, N. T. *et al.* A Weighted and Directed Interareal Connectivity Matrix for Macaque Cerebral Cortex. *Cereb. Cortex* https://doi.org/10.1093/cercor/bhs270 (2012) doi:10.1093/cercor/bhs270.

[14] Bastos, A. M. *et al.* Canonical microcircuits for predictive coding. *Neuron* 76, 695–711 (2012).

[15] Adams, R. A. Bayesian Inference, Predictive Coding, and Computational Models of Psychosis. in *Computational Psychiatry* 175–195 (Elsevier, 2018). doi:10.1016/B978-0-12-809825-7.00007-9.

[16] Sterzer, P. *et al.* The Predictive Coding Account of Psychosis. *Biological Psychiatry* 84, 634–643 (2018).

[17] Seth, A. K., Suzuki, K. & Critchley, H. D. An interoceptive predictive coding model of conscious presence. *Front. Psychol* 3, 395 (2012).

[18] Miłkowski, M. & Litwin, P. Testable or bust: theoretical lessons for predictive processing. *Synthese* 200, 462 (2022).

[19] Bowman, H., Collins, D. J., Nayak, A. K. & Cruse, D. Is predictive coding falsifiable? *Neuroscience & Biobehavioral Reviews* 154, 105404 (2023).

[20] Clark, A. Whatever next? Predictive brains, situated agents, and the future of cognitive science. *Behavioral and Brain Sciences* 36, 181–204 (2013).

[21] Hohwy, J. *The Predictive Mind*. (Oxford University Press, Oxford, 2013).

[22] Arnal, L. H. & Giraud, A.-L. Cortical oscillations and sensory predictions. *Trends Cogn. Sci. (Regul. Ed.)* 16, 390–398 (2012).

[23] Keller, G. B. & Mrsic-Flogel, T. D. Predictive Processing: A Canonical Cortical Computation. *Neuron* 100, 424–435 (2018).

[24] Bastos, A. M., Lundqvist, M., Waite, A. S., Kopell, N. & Miller, E. K. Layer and rhythm specificity for predictive routing. *Proc Natl Acad Sci U S A* 117, 31459–31469 (2020).

[25] Chao, Z. C., Huang, Y. T. & Wu, C.-T. A quantitative model reveals a frequency ordering of prediction and prediction-error signals in the human brain. *Commun Biol* 5, 1–18 (2022).

[26] Chao, Z. C., Takaura, K., Wang, L., Fujii, N. & Dehaene, S. Large-Scale Cortical Networks for Hierarchical Prediction and Prediction Error in the Primate Brain. *Neuron* 100, 1252-1266.e3 (2018).

[27] Garrido, M. I., Kilner, J. M., Kiebel, S. J. & Friston, K. J. Evoked brain responses are generated by feedback loops. *Proc. Natl. Acad. Sci. U.S.A.* 104, 20961–20966 (2007).

[28] Kok, P., Rahnev, D., Jehee, J. F. M., Lau, H. C. & de Lange, F. P. Attention Reverses the Effect of Prediction in Silencing Sensory Signals. *Cerebral Cortex* 22, 2197–2206 (2011).

[29] Bekinschtein, T. A. *et al.* Neural signature of the conscious processing of auditory regularities. *Proceedings of the National Academy of Sciences* 106, 1672–1677 (2009).

[30] Kok, P., Bains, L. J., van Mourik, T., Norris, D. G. & de Lange, F. P. Selective Activation of the Deep Layers of the Human Primary Visual Cortex by Top-Down Feedback. *Current Biology* 26, 371–376 (2016).

[31] Kok, P., Jehee, J. F. M. & de Lange, F. P. Less is more: expectation sharpens representations in the primary visual cortex. *Neuron* 75, 265–270 (2012).

[32] Thomas, E. R. *et al.* Predictions and errors are distinctly represented across V1 layers. *Current Biology* 34, 2265-2271.e4 (2024).

[33] Xiong, Y. (Sophy) *et al.* Propofol-mediated loss of consciousness disrupts predictive routing and local field phase modulation of neural activity. *Proc. Natl. Acad. Sci. U.S.A.* 121, e2315160121 (2024).

[34] Westerberg, J. A. *et al.* Hierarchical substrates of prediction in visual cortical spiking. Preprint at https://doi.org/10.1101/2024.10.02.616378 (2024).

[35] Zmarz, P. & Keller, G. B. Mismatch Receptive Fields in Mouse Visual Cortex. *Neuron* 92, 766–772 (2016).

[36] Bastos, G. *et al.* Top-down input modulates visual context processing through an interneuron-specific circuit. *Cell Reports* 42, 113133 (2023).

[37] Hamm, J. P., Shymkiv, Y., Han, S., Yang, W. & Yuste, R. Cortical ensembles selective for context. *Proc Natl Acad Sci U S A* 118, e2026179118 (2021).

[38] Findling, C. *et al.* Brain-wide representations of prior information in mouse decision-making. *Nature* 645, 192–200 (2025).

[39] Rodriguez, N. Y., Ahuja, A., Basu, D., McKim, T. H. & Desrochers, T. M. Different Subregions of Monkey Lateral Prefrontal Cortex Respond to Abstract Sequences and Their Components. *J. Neurosci.* 44, (2024).

[40] Aizenbud, I. *et al.* Neural mechanisms of predictive processing: a collaborative community experiment through the OpenScope program. Preprint at https://doi.org/10.48550/arXiv.2504.09614 (2026).

[41] Mulvey A.G. *et al.* Deficits in GABAergic Inhibitory Interneurons in Schizophrenia are Specific to Cell type, Cortical layer, and Brain Region. in (Montreal, Canada, 2024).

[42] Cao, C. *et al.* Automation of Systematic Reviews with Large Language Models. 2025.06.13.25329541 Preprint at https://doi.org/10.1101/2025.06.13.25329541 (2026).

[43] Li, L., Mathrani, A. & Susnjak, T. Transforming Evidence Synthesis: A Systematic Review of the Evolution of Automated Meta-Analysis in the Age of AI. Preprint at https://doi.org/10.48550/arXiv.2504.20113 (2025).

[44] Song, Z. *et al.* Evaluating Large Language Models in Scientific Discovery. *arXiv.org* https://arxiv.org/abs/2512.15567v1 (2025).

[45] Lu, C. *et al.* Towards end-to-end automation of AI research. *Nature* 651, 914–919 (2026).

[46] Li, J., Zhang, Q., Yu, Y., Fu, Q. & Ye, D. More Agents Is All You Need. Preprint at https://doi.org/10.48550/arXiv.2402.05120 (2024).

[47] Luo, Z., Kasirzadeh, A. & Shah, N. B. The More You Automate, the Less You See: Hidden Pitfalls of AI Scientist Systems. Preprint at https://doi.org/10.48550/arXiv.2509.08713 (2025).

[48] Aygün, E. *et al.* An AI system to help scientists write expert-level empirical software. *Nature* https://doi.org/10.1038/s41586-026-10658-6 (2026) doi:10.1038/s41586-026-10658-6.

[49] Gabhart, K. M., Xiong, Y. (Sophy) & Bastos, A. M. Predictive coding: a more cognitive process than we thought? *Trends in Cognitive Sciences* 0, (2025).

[50] Solomon, S. S., Tang, H., Sussman, E. & Kohn, A. Limited Evidence for Sensory Prediction Error Responses in Visual Cortex of Macaques and Humans. *Cerebral Cortex* 31, 3136–3152 (2021).

[51] Nourski, K. V. *et al.* Auditory Predictive Coding across Awareness States under Anesthesia: An Intracranial Electrophysiology Study. *J. Neurosci.* 38, 8441–8452 (2018).

[52] Wacongne, C. *et al.* Evidence for a hierarchy of predictions and prediction errors in human cortex. *Proceedings of the National Academy of Sciences* 108, 20754–20759 (2011).

[53] Arcuschin, I., Chanin, D., Garriga-Alonso, A. & Camburu, O.-M. Biases in the Blind Spot: Detecting What LLMs Fail to Mention. *arXiv.org* https://arxiv.org/abs/2602.10117v4 (2026).

[54] Beatty, D., Masanthia, K., Kaphol, T. & Sethi, N. Revealing Hidden Bias in AI: Lessons from Large Language Models. *arXiv.org* https://arxiv.org/abs/2410.16927v1 (2024).

[55] Andrej. karpathy/llm-council. (2026).

[56] Yang, H. *et al.* Large Language Model Synergy for Ensemble Learning in Medical Question Answering: Design and Evaluation Study. *Journal of Medical Internet Research* 27, e70080 (2025).

[57] Li, Z., Tian, H., Luo, L., Cao, Y. & Luo, P. DeepRead: Document Structure-Aware Reasoning to Enhance Agentic Search. Preprint at https://doi.org/10.48550/arXiv.2602.05014 (2026).

[58] White, J. *et al.* A Prompt Pattern Catalog to Enhance Prompt Engineering with ChatGPT. Preprint at https://doi.org/10.48550/arXiv.2302.11382 (2023).

[59] Rao, R. P. N. & Ballard, D. H. Predictive coding in the visual cortex: a functional interpretation of some extra-classical receptive-field effects. *Nat Neurosci* 2, 79–87 (1999).

[60] Attinger, A., Wang, B. & Keller, G. B. Visuomotor Coupling Shapes the Functional Development of Mouse Visual Cortex. *Cell* 169, 1291-1302.e14 (2017).

[61] Bakhtiari, S., Mineault, P., Lillicrap, T., Pack, C. & Richards, B. The functional specialization of visual cortex emerges from training parallel pathways with self-supervised predictive learning. in *Advances in Neural Information Processing Systems* vol. 34 25164–25178 (Curran Associates, Inc., 2021).

[62] Bekinschtein, T. A. *et al.* Neural signature of the conscious processing of auditory regularities. *Proc. Natl. Acad. Sci. U.S.A.* 106, 1672–1677 (2009).

[63] Furutachi, S., Franklin, A. D., Aldea, A. M., Mrsic-Flogel, T. D. & Hofer, S. B. Cooperative thalamocortical circuit mechanism for sensory prediction errors. *Nature* 633, 398–406 (2024).

[64] Garrett, M. *et al.* Experience shapes activity dynamics and stimulus coding of VIP inhibitory cells. *eLife* 9, e50340 (2020).

[65] Greedy, W., Zhu, H. W., Pemberton, J., Mellor, J. & Ponte Costa, R. Single-phase deep learning in cortico-cortical networks. *Advances in Neural Information Processing Systems* 35, 24213–24225 (2022).

[66] Hertäg, L. & Sprekeler, H. Learning prediction error neurons in a canonical interneuron circuit. *eLife* 9, e57541 (2020).

[67] Jiang, L. P. & Rao, R. P. Dynamic predictive coding: A model of hierarchical sequence learning and prediction in the neocortex. *PLOS Computational Biology* 20, e1011801 (2024).

[68] Keller, G. B., Bonhoeffer, T. & Hübener, M. Sensorimotor mismatch signals in primary visual cortex of the behaving mouse. *Neuron* 74, 809–815 (2012).

[69] Kiebel, S. J., Daunizeau, J. & Friston, K. J. A Hierarchy of Time-Scales and the Brain. *PLoS Comput Biol* 4, e1000209 (2008).

[70] B. Lao-Rodríguez, A. *et al.* Neuronal responses to omitted tones in the auditory brain: A neuronal correlate for predictive coding. *Science Advances* https://doi.org/10.1126/sciadv.abq8657 (2023) doi:10.1126/sciadv.abq8657.

[71] Lee, K., Pennartz, C. M. A. & Mejias, J. F. Cortical networks with multiple interneuron types generate oscillatory patterns during predictive coding. *PLOS Computational Biology* 21, e1013469 (2025).

[72] Mikulasch, F. A., Rudelt, L., Wibral, M. & Priesemann, V. Where is the error? Hierarchical predictive coding through dendritic error computation. *Trends in Neurosciences* 46, 45–59 (2023).

[73] Nejad, K. K., Anastasiades, P., Hertäg, L. & Costa, R. P. Self-supervised predictive learning accounts for cortical layer-specificity. *Nature Communications* 16, 6178 (2025).

[74] Payeur, A., Guerguiev, J., Zenke, F., Richards, B. A. & Naud, R. Burst-dependent synaptic plasticity can coordinate learning in hierarchical circuits. *Nat Neurosci* 24, 1010–1019 (2021).

[75] Rao, R. P. N. A sensory–motor theory of the neocortex. *Nat Neurosci* 27, 1221–1235 (2024).

[76] Sacramento, J., Ponte Costa, R., Bengio, Y. & Senn, W. Dendritic cortical microcircuits approximate the backpropagation algorithm. in *Advances in Neural Information Processing Systems* vol. 31 (Curran Associates, Inc., 2018).

[77] Spratling, M. W. Reconciling predictive coding and biased competition models of cortical function. *Front. Comput. Neurosci* 2, 4, (2008).

[78] Spratling, M. W. Predictive coding as a model of response properties in cortical area V1. *J. Neurosci. Off. J. Soc. Neurosci* 30, 3531–3543 (2010).

[79] Van Derveer, A. B. *et al.* A Role for Somatostatin-Positive Interneurons in Neuro-Oscillatory and Information Processing Deficits in Schizophrenia. *Schizophrenia Bulletin* 47, 1385–1398 (2021).

[80] Wacongne, C., Changeux, J.-P. & Dehaene, S. A Neuronal Model of Predictive Coding Accounting for the Mismatch Negativity. *J. Neurosci.* 32, 3665–3678 (2012).

[81] Yamins, D. L. *et al.* Performance-optimized hierarchical models predict neural responses in higher visual cortex. *Proceedings of the national academy of sciences* 111, 8619–8624 (2014).

[82] Liu, Y. *et al.* G-Eval: NLG Evaluation using Gpt-4 with Better Human Alignment. in *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing* 2511–2522 (Association for Computational Linguistics, Singapore, 2023). doi:10.18653/v1/2023.emnlp-main.153.

[83] Hannun, A., Digani, J., Katharopoulos, A. & Collobert, R. ml-explore: Efficient and flexible machine learning on Apple silicon. *GitHub* https://github.com/ml-explore (2026).

[84] Nguyen, M. N. *et al.* Turning Up the Heat: Min-p Sampling for Creative and Coherent LLM Outputs. Preprint at https://doi.org/10.48550/arXiv.2407.01082 (2025).

[85] Rouhani, B. D. *et al.* Microscaling Data Formats for Deep Learning. Preprint at https://doi.org/10.48550/ARXIV.2310.10537 (2023).

[86] Ghareeb, A. E. *et al.* A multi-agent system for automating scientific discovery. *Nature* https://doi.org/10.1038/s41586-026-10652-y (2026) doi:10.1038/s41586-026-10652-y.

[87] Gottweis, J. *et al.* Accelerating scientific discovery with Co-Scientist. *Nature* https://doi.org/10.1038/s41586-026-10644-y (2026) doi:10.1038/s41586-026-10644-y.

Author contributions

H.N. designed the pipeline, implemented the analyses, curated the ontology, and drafted the manuscript. A.M., J.S.-S., and A.M.B. contributed to conceptual framing, interpretation, supervision, and manuscript revision. All authors reviewed the manuscript.

Funding declaration

This research was supported by funding awarded to André M. Bastos from the National Institute of Mental Health R00MH116100 and from the National Science Foundation Career grant 2449277.

Competing interests

The authors declare no competing interests.

Ethics declarations

This study analyzed published literature and model-generated outputs only. No new human participant data, human tissue, live vertebrate animal experiments, or clinical intervention data were collected.

Additional information

Large language models were used as scoring agents within the described local evaluation pipeline. They were not assigned authorship. The authors are responsible for the design, analysis, interpretation, and submitted text.

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- 2026mllmsupp.pdf